Appendix: Proof, Simulation Details, and Supplementary Contexts

Shicheng Liu & Minghui Zhu
School of Electrical Engineering and Computer Science
Pennsylvania State University
University Park, PA 16802, USA
{sfl5539,muz16}@psu.edu

This appendix consists of three sections: section 8 provides some basic notions and notations that will be used in the proof, section 9 presents the proofs of all the lemmas and theorem in the paper, and section 10 gives the simulation details. At last, we provide some extra explanations to help better understand the paper.

8 Notions and notations

We define the concatenated reward feature vector as $\phi_r \triangleq [(\phi_r^{[1]})^\top, \cdots, (\phi_r^{[N_E]})^\top]^\top$ and the augmented reward feature under ω_c as $\phi_{r,\omega_c} \triangleq [\phi_r^\top, c_{\omega_c}]^\top$. The augmented reward feature expectation starting from state-action pair (s,a) under policy π is $\mu^\pi_{r,\omega_c}(s,a) \triangleq \phi_{r,\omega_c}(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a)\mu^\pi_{r,\omega_c}(s')ds'$ and the augmented reward feature expectation starting from state s is $\mu^\pi_{r,\omega_c}(s) \triangleq \int_{a\in\mathcal{A}} \pi(a|s)\mu^\pi_{r,\omega_c}(s,a)da$. Moreover, we denote $\mu_{r,\omega_c}(\pi) \triangleq [(\mu_r(\pi))^\top, J_{\omega_c}(\pi)]^\top$. The empirical augmented reward feature expectation vector is defined as $\hat{\mu}_{r,\omega_c} \triangleq [\hat{\mu}_r^\top, \hat{b}_{\omega_c}]^\top$ and its estimate from learner v is defined as $\hat{\mu}^{[v]}_{r,\omega_c} \triangleq [(\hat{\mu}_r^{[v]})^\top, \hat{b}^{[v]}_{\omega_c}]^\top$. For a given vector $\bar{\eta}$, we define the augmented Q-function as $Q^\pi_{\bar{\eta},\omega_c}(s,a) = \bar{\eta}^\top \mu^\pi_{r,\omega_c}(s,a)$ and the augmented value-function as $V^\pi_{\bar{\eta},\omega_c}(s) = \bar{\eta}^\top \mu^\pi_{r,\omega_c}(s)$. The concatenated cost feature expectation starting from state-action pair (s,a) under policy π is $\mu^\pi_c(s,a) \triangleq \phi_c(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a)\mu^\pi_c(s')ds'$ and the concatenated cost feature expectation starting from state s is $\mu^\pi_c(s) \triangleq \int_{a\in\mathcal{A}} \pi(a|s)\mu^\pi_c(s,a)da$. We denote $\mu_c(\pi) \triangleq E^\pi_{S,A}[\sum_{t=0}^\infty \gamma^t \phi_c(S_t,A_t)]$. The empirical concatenated cost feature expectation is $\hat{\mu}_c \triangleq \frac{1}{m}\sum_{j=1}^m \sum_{t=0}^\infty \gamma^t \phi_c(s_t^j, a_t^j)$. The causal entropy staring from (s,a) under policy π is $H^\pi(s,a) = -\ln \pi(a|s) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a)H^\pi(s')ds'$ and the casual entropy from state s is $H^\pi(s) = \int_{a\in\mathcal{A}} \pi(a|s)H^\pi(a|s)da$. We define the state-action visitation frequency as $\psi^\pi(s,a) \triangleq E^\pi[\sum_{t=0}^\infty \gamma^t \mathbb{1}\{S_t = s\}\mathbb{1}\{A_t = a\}]$ and state visitation frequency as $\psi^\pi(s) \triangleq E^\pi[\sum_{t=0}^\infty \gamma^t \mathbb{1}\{S_t = s\}]$, where $\mathbb{1}\{\cdot\}$ is the indicator function. It is obvious that $\psi^\pi(s) \leq \frac{1}{1-\gamma}$ for any $s \in \mathcal{S}$ and $\int_{s\in\mathcal{S}} \psi^\pi(s)ds = \frac{1}{1-\gamma}$. We use $\mathbf{0}_{m\times n}$ to denote an $m \times n$ matrix whose entries are all zero.

Lemma 4. *For any $(s,a) \in \mathcal{S}\times\mathcal{A}$, any $\omega_c \in \Omega_c$, and any π, $||\mu^\pi_{r,\omega_c}(s)||$, $||\mu^\pi_{r,\omega_c}(s,a)||$, $||\mu^\pi_c(s)||$, and $||\mu^\pi_c(s,a)||$ are bounded.*

Proof. We know that $\mu^\pi_{r,\omega_c}(s,a) = \phi_{r,\omega_c}(s,a) + E^\pi_{S,A}[\sum_{t=1}^\infty \gamma^t \phi_{r,\omega_c}(S_t,A_t)|S_0 = s, A_0 = a]$. Since $||\phi_{r,\omega_c}(s,a)|| \leq \sqrt{\sum_{i=1}^{N_E}(l_r^{[i]}d_1^2) + (\sum_{i=1}^{N_E} l_c^{[i]}d_2)^2)}$, we know that $||\mu^\pi_{r,\omega_c}(s,a)|| \leq \frac{1}{1-\gamma}\sqrt{\sum_{i=1}^{N_E}(l_r^{[i]}d_1^2) + (\sum_{i=1}^{N_E} l_c^{[i]}d_2)^2)}$. Because $\mu^\pi_{r,\omega_c}(s) = \int_{a\in\mathcal{A}} \pi(a|s)\mu^\pi_{r,\omega_c}(s,a)da$, we can see that $||\mu^\pi_{r,\omega_c}(s)||$ is bounded given that $||\mu^\pi_{r,\omega_c}(s,a)||$ is bounded. Analogously, $||\mu^\pi_c(s)||$ and $||\mu^\pi_c(s,a)||$ are also bounded. □

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

Constrained soft Bellman policy. We provide the formula of the constrained soft Bellman policy under continuous state-action space and the policy can be approximated through soft Q learning [1]:

$$\pi_{\eta;\omega_c}(a|s) = \frac{\exp(Q^{\text{soft}}_{\eta;\omega_c}(s,a))}{\exp(V^{\text{soft}}_{\eta;\omega_c}(s))}, \qquad V^{\text{soft}}_{\eta;\omega_c}(s) = \ln(\int_{a\in\mathcal{A}} \exp(Q^{\text{soft}}_{\eta;\omega_c}(s,a))\, da),$$
$$Q^{\text{soft}}_{\eta;\omega_c}(s,a) = \sum_{i=1}^{N_E} (\omega_r^{[i]})^\top \phi_r^{[i]}(s,a) + \lambda c_{\omega_c}(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a) V^{\text{soft}}_{\eta;\omega_c}(s')\, ds'.$$

It is obvious that the constrained soft Bellman policy is continuous in (ω_c, η) as it is a composition of continuous functions of (ω_c, η).

9 Proof

This section focuses on the continuous state-action space and has seven subsections. Subsection 9.1 includes some preliminary results for the later subsections, subsection 9.2 presents the proof of Lemma 1, subsection 9.3 provides some intermediate results for the remaining subsections, subsection 9.4 proves Lemma 3, subsection 9.5 presents the derivation of LGA, subsection 9.6 proves Lemma 2, and subsection 9.7 presents the proof of the theorem.

9.1 Preliminary results

In this section, we prove two lemmas which serve as building blocks for the remaining subsections.

Lemma 5. *The gradient* $\nabla_\eta \ln \pi_{\eta;\omega_c}(a|s) = \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a) - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s)$.

Proof. Define $Z_{s,\eta;\omega_c} \triangleq \exp(V^{\text{soft}}_{\eta;\omega_c}(s))$ and $Z_{a|s,\eta;\omega_c} \triangleq \exp(Q^{\text{soft}}_{\eta;\omega_c}(s,a))$, therefore $Z_{a|s,\eta;\omega_c}$ is smooth in η because it is a composition of logarithmic, exponential and linear functions of η. From Leibniz integral rule, we know that $\nabla_\eta \int_a Z_{a|s,\eta;\omega_c} da = \int_a \nabla_\eta Z_{a|s,\eta;\omega_c} da$. Therefore,

$$\nabla_\eta \ln Z_{s,\eta;\omega_c} = \frac{\int_{a\in\mathcal{A}} \nabla_\eta Z_{a|s,\eta;\omega_c} da}{Z_{s,\eta;\omega_c}} = \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s) \Big[\phi_{r,\omega_c}(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a) \cdot$$
$$\nabla_\eta \ln Z_{s',\eta;\omega_c} ds'\Big] da,$$
$$= \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s) \Big\{\phi_{r,\omega_c}(s,a) + \gamma \int_{s'\in\mathcal{S}} P(s'|s,a) \int_{a'\in\mathcal{A}} \pi_{\eta;\omega_c}(a'|s') \Big[\phi_{r,\omega_c}(s',a')$$
$$+ \gamma \int_{s''\in\mathcal{S}} P(s''|s',a') \nabla_\eta \ln Z_{s'',\eta;\omega_c} ds''\Big] da' ds'\Big\} da.$$

Continuing the expansion, we can get:

$$\nabla_\eta \ln Z_{s,\eta;\omega_c} = E^{\pi_{\eta;\omega_c}}_{S,A}[\sum_{t=0}^{\infty} \gamma^t \phi_{r,\omega_c}(S_t, A_t) | S_0 = s] = \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s),$$
$$\nabla_\eta \ln Z_{a|s,\eta;\omega_c} = \phi_{r,\omega_c}(s,a) + E^{\pi_{\eta;\omega_c}}_{S,A}[\sum_{t=1}^{\infty} \gamma^t \phi_{r,\omega_c}(S_t, A_t) | S_0 = s, A_0 = a] = \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a).$$

Therefore, $\nabla_\eta \ln \pi_{\eta;\omega_c}(a|s) = \nabla_\eta \ln Z_{a|s,\eta;\omega_c} - \nabla_\eta \ln Z_{s,\eta;\omega_c} = \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a) - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s)$. □

Lemma 6. *The gradients* $\nabla_\eta \mu_{r,\omega_c}(\pi_{\eta;\omega_c}) = \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \nabla_\eta \pi_{\eta;\omega_c}(a|s) (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a))^\top da\, ds$, $\nabla_\eta H(\pi_{\eta;\omega_c}) = \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s) \nabla_\eta \ln \pi_{\eta;\omega_c}(a|s) (H^{\pi_{\eta;\omega_c}}(s,a) - 1) da ds$.

Proof.

$$\nabla_\eta \mu_{r,\omega_c}(\pi_{\eta;\omega_c}) = \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \Big[\nabla_\eta \pi_{\eta;\omega_c}(a_0|s_0) (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s_0,a_0))^\top$$

$$+ \pi_{\eta;\omega_c}(a_0|s_0)\nabla_\eta \mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_0,a_0)\Big] da_0 ds_0,$$

$$= \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \Big[\nabla_\eta \pi_{\eta;\omega_c}(a_0|s_0)(\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_0,a_0))^\top + \pi_{\eta;\omega_c}(a_0|s_0)\nabla_\eta\big(\phi_{r,\omega_c}(s_0,a_0)$$

$$+ \gamma \int_{s_1\in\mathcal{S}} P(s_1|s_0,a_0)\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_1)ds_1\big)\Big] da_0 ds_0,$$

$$= \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \Big[\nabla_\eta \pi_{\eta;\omega_c}(a_0|s_0)(\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_0,a_0))^\top + \pi_{\eta;\omega_c}(a_0|s_0)\gamma \int_{s_1\in\mathcal{S}} P(s_1|s_0,a_0)\cdot$$

$$\nabla_\eta \mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_1)ds_1\Big] da_0 ds_0,$$

$$= \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \Big\{\nabla_\eta \pi_{\eta;\omega_c}(a_0|s_0)(\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_0,a_0))^\top + \gamma\pi_{\eta;\omega_c}(a_0|s_0) \int_{s_1\in\mathcal{S}} P(s_1|s_0,a_0)\cdot$$

$$\int_{a_1\in\mathcal{A}} \Big[\nabla_\eta \pi_{\eta;\omega_c}(a_1|s_1)(\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_1,a_1))^\top + \pi_{\eta;\omega_c}(a_1|s_1) \int_{s_2\in\mathcal{S}} P(s_2|s_1,a_1)\nabla_\eta \mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s_2)ds_2\Big]$$

$$da_1 ds_1\Big\} da_0 ds_0.$$

Keep the expansion and we can get:

$$\begin{aligned}\nabla_\eta \mu_{r,\omega_c}(\pi_{\eta;\omega_c}) &= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \nabla_\eta \pi_{\eta;\omega_c}(a|s)(\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s,a))^\top da ds,\\ &= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s)\nabla_\eta \ln \pi_{\eta;\omega_c}(a|s)(\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s,a))^\top da ds,\end{aligned}$$

Analogously, we have that

$$\nabla_\eta H(\pi_{\eta;\omega_c}) = \int_{s_0\in\mathcal{S}} P_0(s_0) \int_{a_0\in\mathcal{A}} \Big[\nabla_\eta \pi_{\eta;\omega_c}(a_0|s_0)(H^{\pi_{\eta;\omega_c}}(s_0,a_0))$$

$$+ \pi_{\eta;\omega_c}(a_0|s_0)\nabla_\eta\big(-\ln \pi_{\eta;\omega_c}(a_0|s_0) + \gamma \int_{s_1\in\mathcal{S}} P(s_1|s_0,a_0)H^{\pi_{\eta;\omega_c}}(s_1)ds_1\big)\Big] da_0 ds_0.$$

Keep the expansion, we have:

$$\nabla_\eta H(\pi_{\eta;\omega_c}) = \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s)\nabla_\eta \ln \pi_{\eta;\omega_c}(a|s)(H^{\pi_{\eta;\omega_c}}(s,a) - 1) da ds.$$

□

9.2 Proof of Lemma 1

The Lagrangian of problem (2) is $L(\pi,\eta;\omega_c) = H(\pi) + \sum_{i=1}^{N_E}(\omega_r^{[i]})^\top(\mu_r^{[i]}(\pi) - \hat{\mu}_r^{[i]}) + \lambda(J_{\omega_c}(\pi) - \hat{b}_{\omega_c})$. To find the maximizer of $\max_{\pi\in\Pi} L(\pi,\eta;\omega_c)$, we remove the constant terms and formulate the following problem:

$$\arg\max_{\pi\in\Pi} \sum_{t=0}^{\infty} \gamma^t E_{S,A}^{\pi}\Big[\sum_{i=1}^{N_E}(\omega_r^{[i]})^\top \phi_r^{[i]}(S_t,A_t) + \lambda c_{\omega_c}(S_t,A_t) - \ln \pi(A_t|S_t)\Big]. \tag{5}$$

From [1], we can see that the continuous constrained soft Bellman policy is the optimal solution of (5) by setting the hyperparameter $\alpha = 1$ in equation (2) in [1].

Let p^* be the optimal value of the primal problem (2) and d^* be the optimal value of the dual problem $\min_\eta G(\eta;\omega_c)$. We know that p^* exists (proof of Lemma 2 in [2]). For any η, $G(\eta;\omega_c)$ is an upper bound of p^* because any optimal solution of the primal problem (2) is a feasible solution of $\max_{\pi\in\Pi} L(\pi,\eta;\omega_c)$, therefore, d^* is finite.

We change the policy π to be time-dependent but force it to be stationary, then

$$\frac{\partial L(\pi,\eta;\omega_c)}{\partial \pi^t(a|s)} = -\gamma^t P(S_t = s)(\ln \pi^t(a|s) + 1)$$

$$+ P(S_t = s)E^{\pi}_{S,A}[\sum_{\tau=t+1}^{\infty} -\gamma^{\tau} \ln \pi^{\tau}(A_\tau|S_\tau)|S_t = s, A_t = a] + P(S_t = s)\cdot$$

$$\Big(\gamma^t \sum_{i=1}^{N_E} (\omega_r^{[i]})^\top \phi_r^{[i]}(s,a) + \lambda\gamma^t c_{\omega_c}(s,a) + E^{\pi}_{S,A}[\sum_{\tau=t+1}^{\infty} \gamma^{\tau} \sum_{i=1}^{N_E} (\omega_r^{[i]})^\top \phi_r^{[i]}(S_\tau, A_\tau)$$

$$+ \lambda c_{\omega_c}(S_\tau, A_\tau)|S_t = s, A_t = a]\Big),$$

$$= \gamma^t P(S_t = s)[H^{\pi}(s,a) - 1 + \eta^\top \mu^{\pi}_{r,\omega_c}(s,a)], \tag{6}$$

where $P(S_t = s)$ is the probability (density) of state s being reached at time t.

From the formula of the constrained soft Bellman policy, we know that $\pi_{\eta;\omega_c}(a|s) > 0$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$, thus $\pi_{\eta;\omega_c}$ is an interior point of Π. Therefore $\frac{\partial L(\pi,\eta;\omega_c)}{\partial \pi(a|s)} = 0$ at $\pi_{\eta;\omega_c}$ and from (6) we have that for any $(s,a) \in \mathcal{S} \times \mathcal{A}$ such that $P(S_t = s) \neq 0$ for at least one time t:

$$H^{\pi_{\eta;\omega_c}}(s,a) - 1 + \eta^\top \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a) = 0. \tag{7}$$

Now, we derive the gradient of $G(\eta;\omega_c) = H(\pi_{\eta;\omega_c}) + \sum_{i=1}^{N_E}(\omega_r^{[i]})^\top(\mu_r^{[i]}(\pi_{\eta;\omega_c}) - \hat{\mu}_r^{[i]}) + \lambda(J_{\omega_c}(\pi_{\eta;\omega_c}) - \hat{b}_{\omega_c})$:

$$\nabla_\eta G(\eta;\omega_c) = \nabla_\eta H(\pi_{\eta;\omega_c}) + \nabla_\eta \mu_{r,\omega_c}(\pi_{\eta;\omega_c})\eta + (\mu_{r,\omega_c}(\pi_{\eta;\omega_c}) - \hat{\mu}_{r,\omega_c}),$$

$$= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s)\nabla_\eta \ln \pi_{\eta;\omega_c}(a|s)(H^{\pi_{\eta;\omega_c}}(s,a) - 1 + (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a))^\top \eta) dads$$

$$+ (\mu_{r,\omega_c}(\pi_{\eta;\omega_c}) - \hat{\mu}_{r,\omega_c}),$$

$$= \mu_{r,\omega_c}(\pi_{\eta;\omega_c}) - \hat{\mu}_{r,\omega_c},$$

where the second equality follows from Lemma 6 and the third equality follows from (7) and the fact that $\psi^{\pi_{\eta;\omega_c}}(s) = 0$ if $P(S_t = s) = 0$ for all time t. Notice that $\mu_{r,\omega_c}(\pi_{\eta;\omega_c}) - \hat{\mu}_{r,\omega_c} = [(\mu_r(\pi_{\eta;\omega_c}) - \hat{\mu}_r)^\top, J_{\omega_c}(\pi_{\eta;\omega_c}) - \hat{b}_{\omega_c}]^\top$. Analogously, we can see that $\nabla_\eta G^{[v]}(\eta;\omega_c) = [(\mu_r(\pi_{\eta;\omega_c}) - \hat{\mu}_r^{[v]})^\top, J_{\omega_c}(\pi_{\eta;\omega_c}) - \hat{b}_{\omega_c}^{[v]}]^\top$.

As $G(\eta;\omega_c)$ attains its minimum at $\eta^*(\omega_c)$, given that $\min_\eta G(\eta;\omega_c)$ is unconstrained, the gradient of $G(\eta;\omega_c)$ with respect to η should be 0 at this point (*i.e.* $\mu_r(\pi_{\eta;\omega_c}) - \hat{\mu}_r = 0$ and $J_{\omega_c}(\pi_{\eta;\omega_c}) - \hat{b}_{\omega_c} = 0$). It implies that $\pi_{\eta;\omega_c}$ is a feasible solution of the primal problem (2). Thus, we have:

$$H(\pi_{\eta^*(\omega_c);\omega_c}) = G(\eta^*(\omega_c);\omega_c) = d^* \geq p^* \geq H(\pi_{\eta^*(\omega_c);\omega_c}),$$

Therefore, we know that $p^* = d^*$ and p^* is obtained at $\pi_{\eta^*(\omega_c);\omega_c}$.

9.3 Intermediate results

In this section, we prove three lemmas which will be used in the remaining subsections.

Lemma 7. *The dual functions $G(\eta;\omega_c)$ and $G^{[v]}(\eta;\omega_c)$ are both strictly convex in η for any nonzero cost weight vector ω_c.*

We know that

$$\nabla^2_{\eta\eta} G(\eta;\omega_c) = \nabla_\eta \mu_{r,\omega_c}(\pi_{\eta;\omega_c}),$$

$$= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s)\nabla_\eta \ln \pi_{\eta;\omega_c}(a|s)(\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a))^\top dads,$$

$$= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s)\Big[\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a) - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s)\Big](\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a))^\top dads,$$

where the second equality follows from Lemma 6 and the last equality follows from Lemma 5.

For any nonzero vector $\bar{\eta}$, we have:

$$\bar{\eta}^\top \nabla^2_{\eta\eta} G(\eta;\omega_c)\bar{\eta} = \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\omega_c,\eta}(a|s)\Big[\eta^\top \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a) - \eta^\top \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s)\Big]$$

$$
\begin{aligned}
&\cdot (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a))^\top \bar{\eta} dads, \\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s) \left[Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,a) - V^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s) \right] (Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,a)) dads, \\
&= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \text{Var}(Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,\cdot)) ds,
\end{aligned}
$$

where $\text{Var}(Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,\cdot))$ is the variance of the augmented Q-function $Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}$ at state s.

We know that $Q^{\pi_{\eta;\omega_c}}_{\bar{\eta};\omega_c}(s,a) = \bar{\eta}^\top \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a)$ and $\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c,j}(s,a)$ is in essence a discounted cumulative sum of $\phi_{r,\omega_c,j}$. Because each component of ϕ_{r,ω_c} can be chosen independently, the variance $\text{Var}(Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,\cdot)) = \sum_j \bar{\eta}_j^2 \text{Var}(\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c,j}(s,\cdot))$. From the formula of the constrained soft Bellman policy, we know that whatever η and ω_c are, $\pi_{\eta;\omega_c}(a|s) > 0$ for any $a \in \mathcal{A}$ at state s. Therefore, the variance $\text{Var}(\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c,j}(s,\cdot))$ is zero only under the situation that $\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c,j}(s,a)$ is a constant almost everywhere over $\mathcal{A}$ and we can choose $\phi_{r,\omega_c,j}$ whose values always vary over a to avoid this. Therefore, $\text{Var}(\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c,j}(s,\cdot)) > 0$ and thus $\text{Var}(Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,\cdot)) > 0$.

Because $\psi^{\pi_{\eta;\omega_c}}(s) \geq 0$ for all $s \in \mathcal{S}$ and $\int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) ds = \frac{1}{1-\gamma}$, the measure of the set where $\psi^{\pi_{\eta;\omega_c}}(s) > 0$ is strictly greater than 0, otherwise $\int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) ds = 0$. Thus, $\int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \text{Var}(Q^{\pi_{\eta;\omega_c}}_{\bar{\eta},\omega_c}(s,\cdot)) ds > 0$ and $\nabla^2_{\eta\eta} G(\eta;\omega_c)$ is positive definite. Analogously, $\nabla^2_{\eta\eta} G^{[v]}(\eta;\omega_c)$ is also positive definite. Therefore, $G(\eta;\omega_c)$ and $G^{[v]}(\eta;\omega_c)$ are both strictly convex for any nonzero ω_c.

Lemma 8. *(i) There is a positive constant $C_{\nabla_\eta G^{[v]}}$ such that for any $\omega_c \in \Omega_c$ and $\eta \in \mathbb{R}^{\sum_{i=1}^{N_E} l_r^{[i]}+1}$, it holds that $||m^{[v]} \nabla_\eta G^{[v]}(\eta;\omega_c)|| \leq C_{\nabla_\eta G^{[v]}}$ and $||m \nabla_\eta G(\eta;\omega_c)|| \leq C_{\nabla_\eta G} \triangleq \sum_{v=1}^{N_L} C_{\nabla_\eta G^{[v]}}$.*
(ii) For any $\omega_c \in \Omega_c$ and $\eta \in \mathbb{R}^{\sum_{i=1}^{N_E}+1}$, $G^{[v]}(\eta;\omega_c)$ and $G(\eta;\omega_c)$ are continuously twice differentiable in (ω_c, η).

Proof. (i) From Lemma 1, we know that $||\nabla_\eta G^{[v]}(\eta;\omega_c)|| = ||\mu_{r,\omega_c}(\pi_{\eta;\omega_c}) - \hat{\mu}^{[v]}_{r,\omega_c}|| \leq ||\mu_{r,\omega_c}(\pi_{\eta;\omega_c})|| + ||\hat{\mu}^{[v]}_{r,\omega_c}||$, thus from Lemma 4, we know that $||m^{[v]} \nabla_\eta G^{[v]}(\eta;\omega_c)|| \leq C_{\nabla_\eta G^{[v]}}$ for some positive constant $C_{\nabla_\eta G^{[v]}}$. Therefore, $||\nabla_\eta m G(\eta;\omega_c)|| = ||\sum_{v=1}^{N_L} m^{[v]} \nabla_\eta G^{[v]}(\eta;\omega_c)|| \leq \sum_{v=1}^{N_L} ||m^{[v]} \nabla_\eta G^{[v]}(\eta;\omega_c)|| \leq \sum_{v=1}^{N_L} C_{\nabla_\eta G^{[v]}} \triangleq C_{\nabla_\eta G}$

(ii) From the proof in Lemma 7, we know that:

$$
\begin{aligned}
\nabla^2_{\eta\eta} G(\eta;\omega_c) &= \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s) \cdot \left[\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a) - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s) \right] (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a))^\top dads, \\
&= E^{\pi_{\eta;\omega_c}}_{S,A} \left[(\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A) - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S)) (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A))^\top \right].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\nabla^3_{\eta\eta\eta} G(\eta;\omega) &= E^{\pi_{\eta;\omega_c}}_{S,A} \left\{ \nabla_\eta \left[(\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A) - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S)) (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A))^\top \right] \right\} \\
&= E^{\pi_{\eta;\omega_c}}_{S,A} \Big[2(\nabla_\eta \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A)) (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A))^\top - (\nabla_\eta \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S)) (\mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A))^\top \\
&\quad - \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S) (\nabla_\eta \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(S,A))^\top \Big],
\end{aligned}
$$

where the formula of $\nabla_\eta \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s)$ can be found in Lemma 6 and the formula of $\nabla_\eta \mu^{\pi_{\eta;\omega_c}}_{r,\omega_c}(s,a)$ can be derived in a similar way to Lemma 6.

With similar process, we can also get the formulas of $\nabla^3_{\omega_c\omega_c\omega_c} G(\eta;\omega_c)$, $\nabla^3_{\eta\omega_c\omega_c} G(\eta;\omega_c)$, $\nabla^3_{\eta\eta\omega_c} G(\eta;\omega_c)$, and $\nabla^3_{\eta\omega_c\eta} G(\eta;\omega_c)$. Therefore, $G(\eta;\omega_c)$ is three-times differentiable. Analogously, $G^{[v]}(\eta;\omega_c)$ is also three-times differentiable. Thus, they are both continuously twice differentiable. □

Lemma 9. *There are positive constants $\tau_{\nabla^2_{\eta\eta}G}$ and $\tau_{\nabla^2_{\eta\eta}G^{[v]}}$ such that $\nabla^2_{\eta\eta}G(\omega_c,\eta^*(\omega_c'))\succeq\tau_{\nabla^2_{\eta\eta}G}I$, $\nabla^2_{\eta\eta}G(\omega_c,\bar{\eta}^{[v]}(\omega_c))\succeq\tau_{\nabla^2_{\eta\eta}G}I$, $\nabla^2_{\eta\eta}G^{[v]}(\omega_c,\eta^*(\omega_c'))\succeq\tau_{\nabla^2_{\eta\eta}G^{[v]}}I$ and $\nabla^2_{\eta\eta}G^{[v]}(\omega_c,\bar{\eta}^{[v]}(\omega_c))\succeq\tau_{\nabla^2_{\eta\eta}G^{[v]}}I$ for all $\omega_c\in\Omega_c\backslash\{0\}$*

Proof. The proof of Lemma 9 is divided into two steps where the first step shows that $\nabla^2_{\eta\eta}G(\omega_c,\eta^*(\omega_c'))\succeq\tau_{\nabla^2_{\eta\eta}G}I$ and the second step shows that $\nabla^2_{\eta\eta}G(\omega_c,\bar{\eta}^{[v]}(\omega_c))\succeq\tau_{\nabla^2_{\eta\eta}G}I$. The part for $\nabla^2_{\eta\eta}G^{[v]}$ can be derived analogously.

Step (i). Lemma 7 shows that $G(\eta;\omega_c)$ is strictly convex in η, thus there is a unique $\eta^*(\omega_c)$. Since the problem $\min_\eta G(\eta;\omega_c)$ is unconstrained and $\nabla_\eta G(\eta^*(\omega_c);\omega_c)=0$ holds for any $\omega_c\in\mathbb{R}^{\sum_{i=1}^{N_E}l_c^{[i]}}$, taking derivative with respect to ω_c on both sides renders

$$\nabla^2_{\eta\omega_c}G(\eta^*(\omega_c);\omega_c)+\nabla^2_{\eta\eta}G(\eta^*(\omega_c);\omega_c)\nabla\eta^*(\omega_c)=0\Rightarrow\nabla\eta^*(\omega_c)=-M(\omega_c,\eta^*(\omega_c))^\top \quad (8)$$

where $M(\omega_c,\eta)\triangleq\nabla^2_{\omega_c\eta}G(\eta;\omega_c)\left[\nabla^2_{\eta\eta}G(\eta;\omega_c)\right]^{-1}$.

Since $\eta^*(\omega_c')$ is differentiable (equation (8)) and $\nabla^2_{\eta\eta}G(\eta;\omega_c)$ is continuous in (ω_c,η) (Lemma 8 (ii)), $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ is continuous in (ω_c,ω_c'). Since Ω_c is compact, $\Omega_c\times\Omega_c$ is compact (Tychonoff's theorem, Theorem 1.9.7 in [3]), then the image of $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ is compact (Theorem 4.14 in [4]). We denote the eigenvalues of $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ by $\lambda_i(\omega_c,\omega_c')$ and we know that $\lambda_i(\omega_c,\omega_c')>0$ for any ω_c and ω_c' because $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ is positive definite and symmetric.

Now, we prove that the image of every $\lambda_i(\omega_c,\omega_c')$ is compact. The characteristic polynomial of $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ is $(\lambda-\lambda_1(\omega_c,\omega_c'))\cdots(\lambda-\lambda_d(\omega_c,\omega_c'))=\lambda^d+p_{d-1}(\omega_c,\omega_c')\lambda^{d-1}+\cdots+p_1(\omega_c,\omega_c')\lambda+p_0(\omega_c,\omega_c')$, where $\lambda_i(\omega_c,\omega_c')$ is the root and $p_i(\omega_c,\omega_c')$ is the coefficient. Each $p_i(\omega_c,\omega_c')$ is in essence a polynomial function of the entries of $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ and the entries of $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ are continuous in (ω_c,ω_c') as $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)$ is continuous in (ω_c,ω_c'), then $p_i(\omega_c,\omega_c')$ is continuous in (ω_c,ω_c') (Theorem 4.7 in [4]). Then $\lambda_i(\omega_c,\omega_c')$ is also continuous in (ω_c,ω_c') (Theorem 3.9.1 in [5]), thus the image of $\lambda_i(\omega_c,\omega_c')$ is compact.

According to Heine-Borel theorem (Theorem 2.41 in [4]), the image of $\lambda_i(\omega_c,\omega_c')$ is closed and bounded. Therefore, $\min\lambda_i(\omega_c,\omega_c')$ exists and belongs to the image of $\lambda_i(\omega_c,\omega_c')$, then $\min\lambda_i(\omega_c,\omega_c')$ is positive. Thus, we can choose a positive number $\tau_{\nabla^2_{\eta\eta}G^*}=\min\{\min\lambda_1(\omega_c,\omega_c'),\cdots,\min\lambda_d(\omega_c,\omega_c')\}$ and $\nabla^2_{\eta\eta}G(\eta^*(\omega_c');\omega_c)\succeq\tau_{\nabla^2_{\eta\eta}G^*}I$.

Step (ii). For the distributed gradient descent in Algorithm 2, we know that (equation (5) in [6])

$$\eta^{[v]}(\omega_c,k)=\sum_{v'=1}^{N_L}[\Phi(k-1,0)]^v_{v'}\eta^{[v']}(0)-\sum_{s=1}^{k-1}\alpha(s-1)\sum_{v'=1}^{N_L}[\Phi(k-1,s)]^v_{v'}\cdot$$
$$m^{[v']}\nabla_\eta G^{[v']}(\eta^{[v']}(\omega_c,s-1);\omega_c)-\alpha(k-1)m^{[v]}\nabla_\eta G^{[v]}(\eta^{[v]}(\omega_c,k-1);\omega_c),$$

where $\Phi(k,s)\triangleq W(s)W(s+1)\cdots W(k)$ is the state transition matrix and $[\Phi(k,s)]^v_{v'}$ is the entry at the v-th row and v'-th column.

We define $y(\omega_c,k)\triangleq\frac{1}{N_L}\sum_{v'=1}^{N_L}\eta^{[v']}(0)-\sum_{s=1}^{k}\alpha(s-1)\sum_{v'=1}^{N_L}\frac{m^{[v']}}{N_L}\nabla_\eta G^{[v']}(\eta^{[v']}(\omega_c,s-1);\omega_c)$, then $y(\omega_c,k+1)=y(\omega_c,k)-\frac{\alpha(k)}{N_L}\sum_{v'=1}^{N_L}m^{[v']}\nabla_\eta G^{[v']}(\eta^{[v']}(\omega_c,k);\omega_c)$.

Following the proof of Lemma 5 (a) and proposition 3 in [6], we can get

$$||y(\omega_c,k+1)-\eta^*(\omega_c)||^2,$$
$$\leq||y(\omega_c,k)-\eta^*(\omega_c)||^2+\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)||$$
$$-\frac{2\alpha(k)m}{N_L}\left[G(y(\omega_c,k);\omega_c)-G(\eta^*(\omega_c);\omega_c)\right]+\frac{(\alpha(k))^2}{N_L}C^2_{\nabla_\eta G}, \quad (9)$$
$$\leq||y(\omega_c,k)-\eta^*(\omega_c)||^2+\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)||+\frac{(\alpha(k))^2}{N_L}C^2_{\nabla_\eta G}, \quad (10)$$

$$||y(\omega_c,k)-\eta^{[v]}(\omega_c,k)|| \le 2\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1}{B_0}}\sum_{v'=1}^{N_L}||\eta^{[v']}(0)||$$
$$+\sum_{s=1}^{k-1}2\alpha(s-1)C_{\nabla_\eta G}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}+\frac{\alpha(k-1)}{N_L}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}}),$$

where $B_0=(N_L-1)B$.

We can find c_η such that $\sum_{v=1}^{N_L}||\eta^{[v]}(0)|| \le c_\eta$. Therefore,

$$\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)|| \le \frac{8\alpha(k)C_{\nabla_\eta G}c_\eta}{N_L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1}{B_0}}$$
$$+\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}\sum_{s=1}^{k-1}2\alpha(s-1)C_{\nabla_\eta G}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}$$
$$+\frac{4\alpha(k)\alpha(k-1)}{N_L^2}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}}),$$
$$=\frac{8\alpha(k)C_{\nabla_\eta G}c_\eta}{N_L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1}{B_0}}+\frac{8\alpha(k)C_{\nabla_\eta G}^2}{N_L}\sum_{s=1}^{k-1}\alpha(s-1)\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}$$
$$+\frac{4\alpha(k)\alpha(k-1)C_{\nabla_\eta G}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L^2}. \tag{11}$$

Notice that the second term in (11) does not exist if $k<2$.

Claim 1. *$\sum_{k=1}^{K}\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)||$ is bounded for any K.*

Proof. Because the upper bound of $\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)||$, i.e., formula (11), is positive, it suffices to show that $\sum_{k=1}^{\infty}\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L}C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)||$ is bounded.

Now, we prove that the summation of each term in (11) is finite one by one.

First, $\sum_{k=1}^{\infty}\frac{8\alpha(k)c_\eta C_{\nabla_\eta G}}{N_L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1}{B_0}} \le \frac{8\bar{\alpha}c_\eta C_{\nabla_\eta G}}{N_L}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}\sum_{k=1}^{\infty}(1-\epsilon^{B_0})^{\frac{k-1}{B_0}} = \frac{8\bar{\alpha}c_\eta C_{\nabla_\eta G}}{N_L}\frac{1+\epsilon^{-B_0}}{(1-\epsilon^{B_0})[1-(1-\epsilon^{B_0})^{\frac{1}{B_0}}]}$ is bounded.

Second,

$$\sum_{k=2}^{\infty}\frac{8\alpha(k)C_{\nabla_\eta G}^2}{N_L}\sum_{s=1}^{k-1}\alpha(s-1)\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}=\frac{8C_{\nabla_\eta G}^2(1+\epsilon^{-B_0})}{N_L(1-\epsilon^{B_0})}\cdot$$
$$\left[\sum_{k=2}^{\infty}\frac{\bar{\alpha}}{k+1}\sum_{s=1}^{k-1}\frac{\bar{\alpha}}{s}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}\right].$$

Let $S_k=\sum_{s=1}^{k-1}\frac{\bar{\alpha}}{s}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}$, then $\frac{k}{\bar{\alpha}}S_{k-1}-\frac{k+1}{\bar{\alpha}}S_k = \frac{k}{\bar{\alpha}}\sum_{s=1}^{k-2}\frac{\bar{\alpha}}{s}(1-\epsilon^{B_0})^{\frac{k-2-s}{B_0}} - \frac{k+1}{\bar{\alpha}}\sum_{s=1}^{k-1}\frac{\bar{\alpha}}{s}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}} = \sum_{s=1}^{k-2}\frac{k}{s}(1-\epsilon^{B_0})^{\frac{k-2-s}{B_0}} - \sum_{s=0}^{k-2}\frac{k+1}{s+1}(1-\epsilon^{B_0})^{\frac{k-2-s}{B_0}} = \sum_{s=1}^{k-2}\frac{k-s}{s(s+1)}(1-\epsilon^{B_0})^{\frac{k-2-s}{B_0}} - (k+1)(1-\epsilon^{B_0})^{\frac{k-2}{B_0}} = \sum_{s=1}^{k-3}\frac{k-s}{s(s+1)}(1-\epsilon^{B_0})^{\frac{k-2-s}{B_0}} + \frac{2}{(k-2)(k-1)} - (k+1)(1-\epsilon^{B_0})^{\frac{k-2}{B_0}}$.

Because $(1-\epsilon^{B_0})^{\frac{k-2}{B_0}}$ decays faster than $\frac{1}{(k-2)(k-1)(k+1)}$, there exists a positive integer $\bar{K}$ such that $\frac{2}{(k-2)(k-1)} - (k+1)(1-\epsilon^{B_0})^{\frac{k-2}{B_0}} > 0$ if $k>\bar{K}$. Therefore, $\frac{k}{\bar{\alpha}}S_{k-1}-\frac{k+1}{\bar{\alpha}}S_k>0$ if $k>\bar{K}$. We can find a positive number M such that $\frac{k+1}{\bar{\alpha}}S_k<M$ for any $k>0$. Then, $\frac{\bar{\alpha}}{k+1}S_k<\frac{\bar{\alpha}^2 M}{(k+1)^2}$ and $\sum_{k=2}^{\infty}\frac{\bar{\alpha}}{k+1}S_k$ is finite.

Third, $\sum_{k=1}^{\infty} \frac{4\alpha(k)\alpha(k-1)C_{\nabla_\eta G}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L^2} = \frac{4C_{\nabla_\eta G}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L^2}\sum_{k=1}^{\infty}\alpha(k)\alpha(k-1) \leq \frac{4C_{\nabla_\eta G}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L^2}\sum_{k=1}^{\infty}\frac{\bar{\alpha}^2}{k^2}$ is bounded. □

Equation (10) shows that $||y(\omega_c,k+1)-\eta^*(\omega_c)||^2 \leq ||y(\omega_c,k)-\eta^*(\omega_c)||^2 + \frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)|| + \frac{(\alpha(k))^2}{N_L}C^2_{\nabla_\eta G}$.

Telescoping from $k=0$ to $K-1$, we have:

$$||y(\omega_c,k)-\eta^*(\omega_c)||^2 \leq ||y(\omega_c,0)-\eta^*(\omega_c)||^2 + \frac{4\alpha^{(0)}}{N_L}\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}}||y(\omega_c,0)-\eta^{[v']}(0)||$$
$$+\sum_{k=1}^{K-1}\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)|| + \sum_{k=0}^{K-1}\frac{(\alpha(k))^2 C^2_{\nabla_\eta G}}{N_L^2} \leq D_{\max},$$

where $D_{\max} = ||y(\omega_c,0)-\eta^*(\omega_c)||^2 + \frac{4\alpha^{(0)}}{N_L}\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}}||y(\omega_c,0)-\eta^{[v']}(0)|| + \sum_{k=1}^{\infty}\frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)|| + \sum_{k=0}^{\infty}\frac{(\alpha(k))^2 C^2_{\nabla_\eta G}}{N_L^2}$ is finite.

Therefore,

$$||\eta^{[v]}(\omega_c,k)-\eta^*(\omega_c)|| \leq ||\eta^{[v]}(\omega_c,k)-y(\omega_c,k)|| + ||y(\omega_c,k)-\eta^*(\omega_c)||,$$
$$\leq 2c_\eta\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1}{B_0}} + \sum_{s=1}^{k-1} 2C_{\nabla_\eta G}\alpha(s-1)\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}}$$
$$+\frac{\alpha(k-1)}{N_L}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}}) + \sqrt{D_{\max}},$$
$$\leq 2c_\eta\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}} + 2C_{\nabla_\eta G}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}\sum_{s=1}^{k-1}\alpha(s-1)(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}} + \frac{\bar{\alpha}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L}$$
$$+\sqrt{D_{\max}},$$
$$\leq 2c_\eta\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}} + 2\bar{\alpha}C_{\nabla_\eta G}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}\sum_{s=1}^{k-1}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}} + \frac{\bar{\alpha}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L},$$
$$+\sqrt{D_{\max}} \leq \bar{D}_{\max},$$

where $\bar{D}_{\max} = 2c_\eta\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}} + 2\bar{\alpha}C_{\nabla_\eta G}\frac{1+\epsilon^{-B_0}}{1-\epsilon^{B_0}}\sum_{s=1}^{\infty}(1-\epsilon^{B_0})^{\frac{k-1-s}{B_0}} + \frac{\bar{\alpha}(C_{\nabla_\eta G}+N_L C_{\nabla_\eta G^{[v]}})}{N_L} + \sqrt{D_{\max}}$ is finite. Because $\eta^*(\omega_c)$ is bounded within a compact set and $||\eta^{[v]}(\omega_c,k)|| \leq ||\eta^*(\omega_c)|| + \bar{D}_{\max}$, we know that $\eta^{[v]}(\omega_c,k)$ is also bounded within a compact set. Then, $||\bar{\eta}^{[v]}(\omega_c)|| = ||\frac{\sum_{j=0}^{K-1}\alpha(k)\eta^{[v]}(\omega_c,k)}{\sum_{j=0}^{K-1}\alpha(k)}|| \leq \max||\eta^{[v]}(\omega_c,j)|| \leq ||\eta^*(\omega_c)|| + \bar{D}_{\max}$ so that $\bar{\eta}^{[v]}(\omega_c)$ is also bounded within a compact set. Following the same idea of η^*, we can find a positive constant $\tau_{G_{\bar{\eta}\bar{\eta}}}$ such that $\nabla^2_{\eta\eta}G(\bar{\eta}^{[v]}(\omega_c);\omega_c) \succeq \tau_{G_{\bar{\eta}\bar{\eta}}}I$ for all $k>0$.

Then, we can find the positive constant $\tau_{\nabla^2_{\eta\eta}G} = \min\{\tau_{\nabla^2_{\eta\eta}G^*}, \tau_{G_{\bar{\eta}\bar{\eta}}}\}$. With similar derivation, we can find the positive constant $\tau_{\nabla^2_{\eta\eta}G^{[v]}}$. □

9.4 Proof of Lemma 3

From Claim 1, we know that there is a positive constant $c_{\max}$ such that $\sum_{k=0}^{K} 2\alpha(k)\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}}||y(\omega_c,k)-\eta^{[v']}(\omega_c,k)|| \leq c_{\max}$ for all $K>0$. From (9), we know that $||y(\omega_c,k+1)-\eta^*(\omega_c)||^2 \leq ||y(\omega_c,k)-\eta^*(\omega_c)||^2 + \frac{4\alpha(k)}{N_L}\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v]}}||y(\omega_c,k)-\eta^{[v]}(\omega_c,k)|| - \frac{2\alpha(k)m}{N_L}\Big[G(y(\omega_c,k);\omega_c)-$

$G(\eta^*(\omega_c);\omega_c)\Big] + \frac{(\alpha(k))^2}{N_L} C^2_{\nabla_\eta G}$, then we have:

$$\sum_{k=0}^{K-1} \frac{2\alpha(k)m}{N_L}\Big[G(y(\omega_c,k);\omega_c) - G(\eta^*(\omega_c);\omega_c)\Big] \leq ||y(\omega_c,0)-\eta^*(\omega_c)||^2 - ||y(\omega_c,k)-\eta^*(\omega_c)||^2$$
$$+ \sum_{k=0}^{K-1} \frac{4\alpha(k)}{N_L} \sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}} ||y(\omega_c,k) - \eta^{[v']}(\omega_c,k)|| + \sum_{k=0}^{K-1} \frac{(\alpha(k))^2 C^2_{\nabla_\eta G}}{N_L^2},$$
$$\Rightarrow \frac{\sum_{k=0}^{K-1}\alpha(k) m G(y(\omega_c,k);\omega_c)}{\sum_{k=0}^{K-1}\alpha(k)} - mG(\eta^*(\omega_c);\omega_c) \leq \frac{N_L}{2\sum_{k=0}^{K-1}\alpha(k)} ||y(\omega_c,0)-\eta^*(\omega_c)||^2$$
$$+ \frac{1}{\sum_{k=0}^{K-1}\alpha(k)} \sum_{k=0}^{K-1} 2\alpha(k) \sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}} ||y(\omega_c,k) - \eta^{[v']}(\omega_c,k)||$$
$$+ \frac{1}{\sum_{k=0}^{K-1}\alpha(k)} \sum_{k=0}^{K-1} \frac{(\alpha(k))^2 C^2_{\nabla_\eta G}}{2N_L},$$
$$\leq \frac{N_L}{2\sum_{k=0}^{K-1}\alpha(k)} ||y(\omega_c,0)-\eta^*(\omega_c)||^2 + \frac{1}{\sum_{k=0}^{K-1}\alpha(k)}(c_{\max} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}^2}{12N_L}),$$
$$\Rightarrow mG(\frac{\sum_{k=0}^{K-1}\alpha(k)y(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)};x) - mG(\eta^*(\omega_c);\omega_c) \leq \frac{\sum_{k=0}^{K-1}\alpha(k) m G(y(\omega_c,k);\omega_c)}{\sum_{k=0}^{K-1}\alpha(k)}$$
$$- mG(\eta^*(\omega_c);\omega_c),$$
$$\leq \frac{N_L}{2\sum_{k=0}^{K-1}\alpha(k)} ||y(\omega_c,0)-\eta^*(\omega_c)||^2 + \frac{1}{\sum_{k=0}^{K-1}\alpha(k)}(c_{\max} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}^2}{12N_L})$$
$$\leq \frac{1}{\log K}\left[\frac{N_L||y(\omega_c,0)-\eta^*(\omega_c)||^2}{2\bar{\alpha}} + c_{\max} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}}{12N_L}\right],$$
$$\Rightarrow ||\frac{\sum_{k=0}^{K-1}\alpha(k)y(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)} - \eta^*(\omega_c)||^2 \leq \frac{2}{\tau_{\nabla^2_{\eta\eta}G}}\Big[G(\frac{\sum_{k=0}^{K-1}\alpha(k)y(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)};\omega_c)$$
$$- G(\eta^*(\omega_c);\omega_c)\Big],$$
$$\leq \frac{2}{m\tau_{\nabla^2_{\eta\eta}G}\log K}\left[\frac{N_L||y(\omega_c,0)-\eta^*(\omega_c)||^2}{2\bar{\alpha}} + c_{\max} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}}{12N_L}\right].$$

Then,

$$||\frac{\sum_{k=0}^{K-1}\alpha(k)\eta^{[v]}(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)} - \eta^*(\omega_c)|| \leq ||\frac{\sum_{k=0}^{K-1}\alpha(k)\eta^{[v]}(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)} - \frac{\sum_{k=0}^{K-1}\alpha(k)y(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)}||$$
$$+ ||\frac{\sum_{k=0}^{K-1}\alpha(k)y(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)} - \eta^*(\omega_c)||,$$
$$\leq \frac{\sum_{k=0}^{K-1}\alpha(k)||\eta^{[v]}(\omega_c,k) - y(\omega_c,k)||}{\sum_{k=0}^{K-1}\alpha(k)} + ||\frac{\sum_{k=0}^{K-1}\alpha(k)y(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)} - \eta^*(\omega_c)||.$$

We know that $\frac{c_{\max}}{2} \geq \sum_{k=0}^{K-1}\alpha(k)\sum_{v'=1}^{N_L} C_{\nabla_\eta G^{[v']}} ||y(\omega_c,k) - \eta^{[v']}(\omega_c,k)|| \geq C_{\nabla_\eta G^{[v]}} \sum_{k=0}^{K-1}$ $\alpha(k)||y(\omega_c,k) - \eta^{[v]}(\omega_c,k)||$, then

$$||\frac{\sum_{k=0}^{K-1}\alpha(k)\eta^{[v]}(\omega_c,k)}{\sum_{k=0}^{K-1}\alpha(k)} - \eta^*(\omega_c)|| \leq \frac{1}{\log K}\frac{c_{\max}}{2C_{\nabla_\eta G^{[v]}}}$$
$$+ \sqrt{\frac{2}{\tau_{\nabla^2_{\eta\eta}G}\log K}\left[\frac{N_L||y(\omega_c,0)-\eta^*(\omega_c)||^2}{2\bar{\alpha}} + c_{\max} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}}{12N_L}\right]},$$

$$= \frac{1}{\log K}\frac{c_{\max}}{2C_{\nabla_\eta G^{[v]}}} + \frac{1}{\sqrt{\log K}}\sqrt{\left[\frac{N_L||y(\omega_c,0)-\eta^*(\omega_c)||^2}{m\tau_{\nabla^2_{\eta\eta}G}\bar{\alpha}} + \frac{2c_{\max}}{m\tau_{\nabla^2_{\eta\eta}G}} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}}{6m\tau_{\nabla^2_{\eta\eta}G}N_L}\right.},$$

$$= \frac{1}{\log K}\frac{c_{\max}}{2C_{\nabla_\eta G^{[v]}}} + \frac{1}{\sqrt{\log K}}\sqrt{\frac{N_L||\frac{1}{N_L}\sum_{v=0}^{N_L} y_v^{(0)}(x)-\eta^*(\omega_c)||^2}{m\tau_{\nabla^2_{\eta\eta}G}\bar{\alpha}} + \frac{2c_{\max}}{m\tau_{\nabla^2_{\eta\eta}G}} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}}{6m\tau_{\nabla^2_{\eta\eta}G}N_L}}.$$

We let $C_1^{[v]} = \frac{c_{\max}}{2C_{\nabla_\eta G^{[v]}}}$ and $C_2^{[v]} = \sqrt{\frac{N_L||\frac{1}{N_L}\sum_{v=0}^{N_L} y_v^{(0)}(x)-\eta^*(\omega_c)||^2}{m\tau_{\nabla^2_{\eta\eta}G}\bar{\alpha}} + \frac{2c_{\max}}{m\tau_{\nabla^2_{\eta\eta}G}} + \frac{C^2_{\nabla_\eta G}\pi^2\bar{\alpha}}{6m\tau_{\nabla^2_{\eta\eta}G}N_L}}$.

9.5 The derivation of learner v's LGA

Using the chain rule, we know that:

$$\begin{aligned}\nabla F(\omega_c,\eta^*(\omega_c)) &= \nabla_{\omega_c}F(\omega_c,\eta^*(\omega_c)) + [(\nabla_\eta F(\omega_c,\eta^*(\omega_c)))^\top \nabla\eta^*(\omega_c)]^\top,\\ &= \nabla_{\omega_c}F(\omega_c,\eta^*(\omega_c)) - M(\omega_c,\eta^*(\omega_c))\nabla_\eta F(\omega_c,\eta^*(\omega_c)),\end{aligned}$$

where $M(\omega_c,\eta)$ is defined in (8).

Lemma 10. *The gradient* $\nabla F(\omega,\eta^*(\omega_c)) = \nabla_{\omega_c}F(\omega_c,\eta^*(\omega_c))$.

Proof. The global likelihood function is:

$$F(\omega_c,\eta) = \sum_{t=0}^{\infty}\gamma^t\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}} mP_{\mathcal{D}}(S_t=s,A_t=a)\ln\pi_{\eta;\omega_c}(a|s)dads,$$

where $P_{\mathcal{D}}(S_t=s,A_t=a)$ is the empirical probability of (s,a) occurring at time t in the demonstrations $\mathcal{D}$:

$$P_{\mathcal{D}}(S_t=s,A_t=a) \triangleq \frac{1}{m}\sum_{j=1}^{m}(\mathbb{1}\{s_t^j=s\}\mathbb{1}\{a_t^j=a\}),$$

From the proof of Lemma 5, we know that $\pi_{\eta;\omega_c}$ can be formulated as:

$$\begin{aligned}&\pi_{\eta;\omega_c}(a|s) = \frac{Z_{a|s,\eta;\omega_c}}{Z_{s,\eta;\omega_c}},\\ &\ln Z_{a|s,\eta;\omega_c} = \sum_{i=1}^{N_E}(\omega_r^{[i]})^\top\phi_r^{[i]}(s,a) + \lambda c_{\omega_c}(s,a) + \gamma\int_{s'\in\mathcal{S}}P(s'|s,a)\ln Z_{s',\eta;\omega_c}ds',\\ &\ln Z_{s,\eta;\omega_c} = \ln\int_{a\in\mathcal{A}}Z_{a|s,\eta;\omega_c}da.\end{aligned}$$

Thus,

$$\begin{aligned}&\nabla_\eta F(\omega_c,\eta) = \sum_{t=0}^{\infty}\gamma^t\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}} mP_{\mathcal{D}}(S_t=s,A_t=a)\nabla_\eta(\ln Z_{a|s,\eta;\omega_c} - \ln Z_{s,\eta;\omega_c})dads,\\ &= \sum_{t=0}^{\infty}\gamma^t\int_{s\in\mathcal{S}}\int_{a\in\mathcal{A}} mP_{\mathcal{D}}(S_t=s,A_t=a)\Big\{\phi_{r,\omega_c}(s,a) + E^{\pi_{\eta;\omega_c}}_{S,A}[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t}\phi_{r,\omega_c}(S_\tau,A_\tau)|\\ &S_t=s,A_t=a] - E^{\pi_{\eta;\omega_c}}_{S,A}[\sum_{\tau=t}^{\infty}\gamma^{\tau-t}\phi_{r,\omega_c}(S_\tau,A_\tau)|S_t=s]\Big\}dads,\end{aligned}$$

where the last inequality follows from Lemma 5. Here,

$$\begin{aligned}&\gamma\int_{s'\in\mathcal{S}}\int_{a'\in\mathcal{A}}P_{\mathcal{D}}(S_{t+1}=s',A_{t+1}=a')E^{\pi_{\eta;\omega_c}}_{S,A}\left[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t-1}\phi_{r,\omega_c}(S_\tau,A_\tau)|S_{t+1}=s'\right]da'ds',\\ &= \gamma\int_{s'\in\mathcal{S}}P_{\mathcal{D}}(S_{t+1}=s')E^{\pi_{\eta;\omega_c}}_{S,A}\left[\sum_{\tau=t+1}^{\infty}\gamma^{\tau-t-1}\phi_{r,\omega_c}(S_\tau,A_\tau)|S_{t+1}=s'\right]ds',\end{aligned}$$

$$= \gamma \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} P_{\mathcal{D}}(S_t = s, A_t = a) \int_{s'\in\mathcal{S}} P(s'|s,a) \cdot$$
$$E_{S,A}^{\pi_{\eta;\omega_c}} \Bigg[\sum_{\tau=t+1}^{\infty} \gamma^{\tau-t-1} \phi_{r,\omega_c}(S_\tau, A_\tau) | S_{t+1} = s' \Bigg] ds' da ds,$$
$$= \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} P_{\mathcal{D}}(S_t = s, A_t = a) E_{S,A}^{\pi_{\eta;\omega_c}} \Bigg[\sum_{\tau=t+1}^{\infty} \gamma^{\tau-t} \phi_{r,\omega_c}(S_\tau, A_\tau) | S_t = s, A_t = a \Bigg] da ds.$$

Therefore,

$$\nabla_\eta F(\omega_c, \eta) = \sum_{t=0}^{\infty} \gamma^t \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} m P_{\mathcal{D}}(S_t = s, A_t = a) \phi_{r,\omega_c}(s,a) da ds$$
$$- \int_{s\in\mathcal{S}} \int_{a\in\mathcal{A}} m P_{\mathcal{D}}(S_0 = s, A_0 = a) E_{S,A}^{\pi_{\eta;\omega_c}} \Bigg[\sum_{t=0}^{\infty} \gamma^t \phi_{r,\omega_c}(S_t, A_t) | S_0 = s \Bigg] da ds,$$
$$= m(\hat{\mu}_{r,\omega_c} - \mu_{r,\omega_c}(\pi_{\eta;\omega_c})).$$

The above derivation shows that $\nabla_\eta F(\omega_c; \eta) = -m \nabla_\eta G(\eta; \omega_c)$. Therefore, we have that $\nabla F(\omega_c, \eta^*(\omega_c)) = \nabla_{\omega_c} F(\omega_c, \eta^*(\omega_c)) - M(\omega_c, \eta^*(\omega_c)) \nabla_\eta F(\omega_c, \eta^*(\omega_c)) = \nabla_{\omega_c} F(\omega_c, \eta^*(\omega_c)) + M(\omega_c, \eta^*(\omega_c)) \nabla_\eta m G(\eta^*(\omega_c); \omega_c) = \nabla_{\omega_c} F(\omega_c, \eta^*(\omega_c))$. □

However, we cannot obtain the gradient $\nabla F(\omega_c, \eta^*(\omega_c))$ because each learner v does not find $\eta^*(\omega_c)$ but its approximation $\bar{\eta}^{[v]}(\omega_c)$. Therefore, we propose a surrogate gradient to approximate $\nabla F(\omega_c, \eta^*(\omega_c))$: $\bar{\nabla} F(\omega_c, \bar{\eta}^{[v]}(\omega_c)) \triangleq \nabla_{\omega_c} F(\omega_c, \bar{\eta}^{[v]}(\omega_c))$.

Notice that this surrogate gradient is global, to decompose it, we propose a local gradient approximation from learner v: $\bar{\nabla} F^{[v]}(\omega_c, \bar{\eta}^{[v]}(\omega_c)) \triangleq \nabla_{\omega_c} F^{[v]}(\omega_c, \bar{\eta}^{[v]}(\omega_c))$. This local approximation aims to approximate $\bar{\nabla} F^{[v]}(\omega_c, \eta^*(\omega_c)) = \nabla_{\omega_c} F^{[v]}(\omega_c, \eta^*(\omega_c))$. Notice that $\nabla F(\omega_c, \eta^*(\omega_c)) = \nabla_{\omega_c} F(\omega_c, \eta^*(\omega_c)) = \sum_{v=1}^{N_L} \nabla_{\omega_c} F^{[v]}(\omega_c, \eta^*(\omega_c)) = \sum_{v=1}^{N_L} \bar{\nabla} F^{[v]}(\omega_c, \eta^*(\omega_c))$.

Similar to the derivation of $\nabla_\eta F(\omega_c, \eta)$, we can get the gradient $\nabla_{\omega_c} F^{[v]}(\omega_c, \eta) = \sum_{\zeta^j \in \mathcal{D}^{[v]}} \sum_{t=0}^{\infty} \gamma^t \phi_c(s_t^j, a_t^j) - m^{[v]} E_{S,A}^{\pi_{\eta;\omega_c}} [\sum_{t=0}^{\infty} \gamma^t \phi_c(S_t, A_t)]$. We still use this approximation when $\omega_c = 0$.

Lemma 11. *The approximation error $||\bar{\nabla} F^{[v]}(\omega_c, \bar{\eta}(\omega_c)) - \bar{\nabla} F^{[v]}(\omega_c, \eta^*(\omega_c))||$ is upper bounded and decreases to zero at the rate of $O(\frac{1}{\sqrt{\log K}})$.*

Proof. At first, we first prove the following claim showing that bounded gradient can imply Lipschitz condition of a function if the domain of the function is convex.

Claim 2. *Suppose a function $f : E \subseteq \mathbb{R}^n \to \mathbb{R}^m$ is differentiable on the convex set E, and $||\nabla f(x)|| \leq M$ for any $x \in E$, where M is a finite positive constant. Then f satisfies the Lipschitz condition, i.e., $||f(b) - f(a)|| \leq M||b - a||$ for any $a, b \in E$.*

Proof. Let $y(t) = (1-t)a + tb$ where $t \in [0,1]$ and $h(t) = f \circ y(t)$. Let $z = f(b) - f(a)$ and define $v(t) = z^\top h(t)$. Then v is real-valued and differentiable on $(0,1)$. By the mean value theorem, there exists $\bar{t} \in (0,1)$ such that $v(1) - v(0) = v'(t) \Rightarrow ||f(b) - f(a)||^2 = (f(b) - f(a))^\top (\nabla_y f(y(t)))^\top (b - a) \leq ||f(b) - f(a)|| \cdot ||\nabla_y f(y(t))|| \cdot ||b - a|| \Rightarrow ||f(b) - f(a)|| \leq M||b - a||$. □

Claim 3. *For any $\omega_c \in \Omega_c$, $\nabla_{\omega_c} F^{[v]}(\omega_c, \eta)$ is Lipschitz continuous (w.r.t. η) with positive constant $L_{\nabla_{\omega_c} F^{[v]}}$.*

Proof. From Claim 2, it suffices to show that $||\nabla_{\omega_c \eta} F^{[v]}(\omega_c, \eta)||$ is bounded. Following similar derivation in the proof of Lemma 10, we can get $\nabla_{\omega_c} F^{[v]}(\omega_c, \eta) = m^{[v]}(\hat{\mu}_c - \lambda \mu_c(\pi_{\eta;\omega_c}))$. Then following the similar idea of the proof in Lemma 7, we can see that

$$\nabla^2_{\omega_c \eta} F^{[v]}(\omega_c, \eta) = -m^{[v]} \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \Bigg[\nabla_\eta \pi_{\eta;\omega_c}(a|s) (\lambda \mu_c^{\pi_{\eta;\omega_c}}(s,a))^\top + \pi_{\eta;\omega_c}(a|s)$$

$$\cdot [\mathbf{0}_{\sum_{i=1}^{N_E} l_c^{[i]} \times \sum_{i=1}^{N_E} l_r^{[i]}}, \mu_c^{\pi_{\eta;\omega_c}}(s,a)]^\top \Big] dads$$
$$= m^{[v]} \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \lambda \pi_{\eta;\omega_c}(a|s) [\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s) - \mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s,a)](\mu_c^{\pi_{\eta;\omega_c}}(s,a))^\top dads$$
$$- m^{[v]} \int_{s\in\mathcal{S}} \psi^{\pi_{\eta;\omega_c}}(s) \int_{a\in\mathcal{A}} \pi_{\eta;\omega_c}(a|s) [\mathbf{0}_{\sum_{i=1}^{N_E} l_c^{[i]} \times \sum_{i=1}^{N_E} l_r^{[i]}}, \mu_c^{\pi_{\eta;\omega_c}}(s,a)]^\top dads.$$

Because $\psi^{\pi_{\eta;\omega_c}}(s)$, $\pi_{\eta;\omega_c}(a|s)$, $\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s)$, $\mu_{r,\omega_c}^{\pi_{\eta;\omega_c}}(s,a)$, $\mu_c^{\pi_{\eta;\omega_c}}(s)$, and $\mu_c^{\pi_{\eta;\omega_c}}(s,a)$ are finite (Lemma 4), and $\int_{s\in\mathcal{S}} ds = \mathcal{S}$ and $\int_{a\in\mathcal{A}} da = \mathcal{A}$ are finite, we know that $||\nabla_{\omega_c\eta} F^{[v]}(\omega_c,\eta)||$ is bounded and thus $L_{\nabla_{\omega_c} F^{[v]}}$ exists. □

Therefore, the approximation error is

$$||\bar{\nabla} F^{[v]}(\omega_c, \bar{\eta}(\omega_c)) - \bar{\nabla} F^{[v]}(\omega_c, \eta^*(\omega_c))|| = ||\nabla_{\omega_c} F^{[v]}(\omega_c, \bar{\eta}(\omega_c)) - \nabla_{\omega_c} F^{[v]}(\omega_c, \eta^*(\omega_c))||,$$
$$\leq L_{\nabla_{\omega_c} F^{[v]}} ||\bar{\eta}(\omega_c) - \eta^*(\omega_c)||.$$

From Lemma 3, we know that $||\bar{\eta}(\omega_c) - \eta^*(\omega_c)||$ decreases to 0 at the rate of $O(\frac{1}{\sqrt{\log K}})$, thus $||\bar{\nabla} F^{[v]}(\omega_c, \bar{\eta}(\omega_c)) - \bar{\nabla} F^{[v]}(\omega_c, \eta^*(\omega_c))||$ decreases to 0 at the rate of $O(\frac{1}{\sqrt{\log K}})$. □

9.6 Proof of Lemma 2

Learner v's LSCA problem at $\omega_c^{[v]}$ is $\arg\max_{\omega_c\Omega_c} \tilde{F}^{[v]}(\omega_c; \omega_c^{[v]})$ where $\tilde{F}^{[v]}(\omega_c; \omega_c^{[v]})$ needs to satisfy the following four conditions (Assumption 3.14 in [7]):
(i) $\nabla \tilde{F}^{[v]}(\omega_c; \omega_c^{[v]}) = N_L \bar{\nabla}^{[v]}$ at $\omega_c^{[v]}$.
(ii) $-\tilde{F}^{[v]}(\omega_c; \omega_c^{[v]})$ is strongly convex in ω_c.
(iii) $\tilde{F}^{[v]}(\omega_c; \omega_c^{[v]})$ is continuously differentiable in ω_c.
(iv) $\nabla \tilde{F}^{[v]}(\omega_c; \omega_c^{[v]})$ is Lipschitz continuous in $\omega_c^{[v]}$.

To satisfy these assumptions, we choose the function $\tilde{F}^{[v]}(\omega_c; \omega_c^{[v]}) = -\frac{1}{2}||\omega_c - \omega_c^{[v]}||^2 + (N_L \bar{\nabla}^{[v]})^\top (\omega_c - \omega_c^{[v]})$.

Then $\tilde{\omega}_c^{[v]} = \arg\max_{\omega_c\in\Omega_c} -\frac{1}{2}||\omega_c - \omega_c^{[v]}||^2 + \left(N_L \bar{\nabla}^{[v]}\right)^\top \left(\omega_c - \omega_c^{[v]}\right)$. Now, we prove that $\tilde{\omega}_c^{[v]} = \text{Project}_{\Omega_c}\left(\omega_c^{[v]} + N_L \bar{\nabla}^{[v]}\right)$. Because $\omega_c \in \Omega_c$, by the property of the projection operator, we know that $\tilde{\omega}_c^{[v]} = \arg\min_{\omega_c\in\Omega_c} ||\omega_c - \omega_c^{[v]} - N_L \bar{\nabla}^{[v]}|| = \arg\min_{\omega_c\in\Omega_c} ||\omega_c - \omega_c^{[v]} - N_L \bar{\nabla}^{[v]}||^2 = \arg\min_{\omega_c\in\Omega_c} ||\omega_c - \omega_c^{[v]}||^2 + ||N_L \bar{\nabla}^{[v]}||^2 - 2\left(N_L \bar{\nabla}^{[v]}\right)^\top \left(\omega_c - \omega_c^{[v]}\right) = \arg\max_{\omega_c\in\omega_c} -\frac{1}{2}||\omega_c - \omega_c^{[v]}||^2 + \left(N_L \bar{\nabla}^{[v]}\right)^\top \left(\omega_c - \omega_c^{[v]}\right)$.

9.7 Proof of Theorem 1

With the the result of LSCA in subsection 9.6, SONATA (Theorem 4 in [8]) can prove that

$$\text{(consensus)} \qquad \lim_{n\to\infty} \max_{v,v'\in\mathcal{V}} ||\omega_c^{[v]}(n) - \omega_c^{[v']}(n)|| = 0,$$

$$\text{(convergence)} \qquad \lim_{n\to\infty} (\sum_{v=1}^{N_L} \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n))))^\top (\omega_c - \omega_c^{[v]}(n)) \leq 0$$

Then,

$$\limsup_{n\to\infty} (\nabla F(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))))^\top (\omega_c - \omega_c^{[v]}(n))$$

$$
\begin{aligned}
&= \limsup_{n\to\infty}(\nabla F(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))) - \sum_{v=1}^{N_L} \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n))))^\top (\omega_c - \omega_c^{[v]}(n)) \\
&+ \limsup_{n\to\infty}(\sum_{v=1}^{N_L} \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n))))^\top (\omega_c - \omega_c^{[v]}(n)), \\
&\leq \limsup_{n\to\infty}(\nabla F(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))) - \sum_{v=1}^{N_L} \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n))))^\top (\omega_c - \omega_c^{[v]}(n)), \\
&\leq \limsup_{n\to\infty} ||\nabla F(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))) - \sum_{v=1}^{N_L} \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n)))|| \cdot ||\omega_c - \omega_c^{[v]}(n)||, \\
&\leq \limsup_{n\to\infty} 2 \sum_{i=1}^{N_E} l_c^{[i]} || \sum_{v=1}^{N_L} (\bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))) - \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n))))|| \\
&\leq \limsup_{n\to\infty} 2 \sum_{i=1}^{N_E} l_c^{[i]} \sum_{v=1}^{N_L} ||\bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))) - \bar{\nabla} F^{[v]}(\omega_c^{[v]}(n), \bar{\eta}^{[v]}(\omega_c^{[v]}(n)))|| \\
&\leq \limsup_{n\to\infty} 2 \sum_{i=1}^{N_E} l_c^{[i]} \sum_{v=1}^{N_L} L_{\nabla_{\omega_c} F^{[v]}} ||\eta^*(\omega_c^{[v]}(n)) - \bar{\eta}^{[v]}(\omega_c^{[v]}(n))||,
\end{aligned}
$$

where the last inequality follows from Claim 3 and $||\eta^*(\omega_c^{[v]}(n)) - \bar{\eta}^{[v]}(\omega_c^{[v]}(n))|| = O(\frac{1}{\sqrt{\log K}})$ (Lemma 3). Then, we have

$$
\limsup_{n\to\infty}(\nabla F(\omega_c^{[v]}(n), \eta^*(\omega_c^{[v]}(n))))^\top (\omega_c - \omega_c^{[v]}(n)) \leq \frac{\bar{M}}{\sqrt{\log K}},
$$

where $\bar{M}$ is well-defined because $l_c^{[i]}$ and $L_{\nabla_{\omega_c}} F^{[v]}$ are both positive constants.

10 Simulation details

In this section, we provide the simulation details. The Python3 code was run on a laptop with one Intel Core i7-9750H 2.60GHz CPU and 16 GB of RAM under Ubuntu 18.04 operating system. Due to the well-known curse of dimensionality, the reinforcement learning (or dynamic programming) of multiple experts is hard to compute. To alleviate this issue, we model the experts as separate MDPs for most of the time and only model them as an MG when they are close to each other.

10.1 Synthetic grid world

The inner communication network for the four learners has two stages: in stage 1, learners 1 and 2 can communicate and learners 3 and 4 can communicate; in stage 2, learners 1 and 4 can communicate and learners 2 and 3 can communicate. Thus the adjacency matrix in stage 1 is $\begin{bmatrix} 0.5 & 0.5 & 0 & 0 \\ 0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 0.5 & 0.5 \end{bmatrix}$ and in stage 2 is $\begin{bmatrix} 0.5 & 0 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \\ 0 & 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0 & 0.5 \end{bmatrix}$.

The outer communication graph also has two stages, as we need to ensure $\bar{N}^{[v]}(n) = \mathcal{N}$, the adjacency matrix in two stages are respectively:
$\begin{bmatrix} 0.49999 & 0.49999 & 0.00001 & 0.00001 \\ 0.49999 & 0.49999 & 0.00001 & 0.00001 \\ 0.00001 & 0.00001 & 0.49999 & 0.49999 \\ 0.00001 & 0.00001 & 0.49999 & 0.49999 \end{bmatrix}$ and $\begin{bmatrix} 0.49999 & 0.00001 & 0.00001 & 0.49999 \\ 0.00001 & 0.49999 & 0.49999 & 0.00001 \\ 0.00001 & 0.49999 & 0.49999 & 0.00001 \\ 0.49999 & 0.00001 & 0.00001 & 0.49999 \end{bmatrix}$.

The discount factor is 0.9. For each $\omega_{c,j}^{[i]}$, we also need a threshold $\omega_{c,\text{th}}$ to judge whether $\mathcal{C}_j^{[i]}$ is identified as a part of the ground truth constraint, i.e., $\mathcal{C}_j^{[i]}$ is identified as a part of the ground truth constraint if $\omega_{c,j}^{[i]} \geq \omega_{c,\text{th}}$ and is not if otherwise. The principle of choosing the value of $\omega_{c,\text{th}}$ is that

the probability of entering $\mathcal{C}_j^{[i]}$ when $\omega_{c,j}^{[i]} = \omega_{c,\text{th}}$ is less than 1% of the probability when $\omega_{c,j}^{[i]} = 0$. In this example, we choose $\omega_{c,\text{th}} = 0.1$.

10.2 Drones motion planning with obstacles

The simulator is built in Gazebo based on a package called hector_quadrotor [9].

In this experiment, there are only two learners who obtain, respectively, four and five pairs of demonstrated trajectories. The inner and outer communication graphs are the same and only have one stage. The adjacency matrix is $\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}$. The discount factor is 1 and the drones fly at most 500 steps and we choose $\omega_{c,\text{th}} = 0.05$.

11 Supplementary contexts

11.1 Intuition of the bi-level formulation

The intuition behind this bi-level formulation is inspired by hyperparameter learning [10] and an IRL literature [11]. Paper [11] interprets IRL as an bi-level problem where the outer level is to learn the reward function and the inner level is to learn the corresponding policy. In our case, we treat the constraint as a hyperparameter in our learned environment and we need to recover the corresponding reward function and policy given current constraint. Therefore, in our problem, the outer level is to learn the constraints and the inner level is to learn the corresponding reward function and policy.

11.2 Derivation of centralized bi-level problem and decomposition to the distributed bi-level problem

The global log likelihood of the global demonstration set $\mathcal{D} = \{\zeta^j\}_{j=1}^m$ is $F(\omega_c) = \sum_{j=1}^m \sum_{t=0}^\infty \gamma^t \ln \pi_{\omega_c}(a_t^j | s_t^j)$. To solve this MLE problem, we need to find a parametric policy model of π_{ω_c}. To find such a model, we use an optimization problem (2) based on MCE scheme. Notice that this optimization problem (2) is parameterized by ω_c, thus its optimal solution is also parameterized by ω_c. Its optimal solution is the policy model we want in the likelihood function $F(\omega_c)$. Following well-established MCE IRL literature [2], we can see that the optimal solution of this optimization problem is the constrained soft Bellman policy $\pi_{\eta^*(\omega_c);\omega_c}$, where $\eta^*(\omega_c)$ is the optimization solution of the problem $\min_\eta G(\eta; \omega_c)$. To fully define the MLE problem, we also need the optimal solution $\eta^*(\omega_c)$ of the problem $\min_\eta G(\eta; \omega_c)$. Therefore, we formulate the MLE problem as a bi-level optimization problem: $\max_{\omega_c} F(\omega_c, \eta^*(\omega_c)) = \sum_{j=1}^m \sum_{t=0}^\infty \gamma^t \ln \pi_{\eta^*(\omega_c);\omega_c}(a_t^j | s_t^j)$ s.t. $\eta^*(\omega_c) = \arg\min_\eta G(\eta; \omega_c)$. Now, we finish the derivation of the centralized bi-level problem.

As no learner knows the global demonstration data $\mathcal{D}$ and no learner can formulate F and G, the centralized bi-level problem cannot be directly solved. Therefore, we need to decompose the centralized problem into an equivalent distributed problem that the learners can solve even if each learner only knows its local demonstration. Therefore, we decompose F into multiple $F^{[v]}$ and G into multiple $G^{[v]}$. In remark 3, we show that the distributed problem (3)-(4) is equivalent to the centralized bi-level problem we derive here.

11.3 Explanation of the theoretical results

In our setting, each distributed learner only has a portion of the global demonstration data set and one goal of our distributed algorithm is that the distributed learners can learn as good as a centralized learner who obtains the global data set even if the distributed learners do not share their local demonstration set. In Lemma 3, it is shown that given a cost feature estimate ω_c, the distributed learners can converge to a consensus and the consensus is the optimal solution of the inner problem, i.e., their consensus is the best solution the centralized learner can achieve. In Theorem 1, it is shown that the distributed learners can achieve consensus on the cost weight ω_c which belongs to the stationary point set of the outer problem. Combining Lemma 3 and Theorem 1, we can see that the distributed learners will converge to the consensus on both the reward function and cost

function where the learned cost weight belongs to the stationary point set of the outer problem and the learned reward weight is the optimal solution of the inner problem.

References

[1] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *International Conference on Machine Learning*, pp. 1352–1361, 2017.

[2] Z. Zhou, M. Bloem, and N. Bambos, "Infinite time horizon maximum causal entropy inverse reinforcement learning," *IEEE Transactions on Automatic Control*, vol. 63, no. 9, pp. 2787–2802, 2017.

[3] V. I. Bogachev and O. G. Smolyanov, *Real and functional analysis*. Springer, 2020.

[4] W. Rudin, *Principles of mathematical analysis*. McGraw-hill New York, 1976.

[5] E. E. Tyrtyshnikov, *A brief introduction to numerical analysis*. Springer Science & Business Media, 1997.

[6] A. Nedic and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[7] G. Scutari and Y. Sun, "Parallel and distributed successive convex approximation methods for big-data optimization," in *Multi-agent Optimization*, pp. 141–308, 2018.

[8] G. Scutari and Y. Sun, "Distributed nonconvex constrained optimization over time-varying digraphs," *Mathematical Programming*, vol. 176, no. 1, pp. 497–544, 2019.

[9] J. Meyer, A. Sendobry, S. Kohlbrecher, U. Klingauf, and O. von Stryk, "Comprehensive simulation of quadrotor uavs using ros and gazebo," in *International Conference on Simulation, Modeling and Programming for Autonomous Robots*, p. to appear, 2012.

[10] K. Ji, J. Yang, and Y. Liang, "Bilevel optimization: Convergence analysis and enhanced design," in *International Conference on Machine Learning*, pp. 4882–4892, 2021.

[11] N. Das, S. Bechtle, T. Davchev, D. Jayaraman, A. Rai, and F. Meier, "Model-based inverse reinforcement learning from visual demonstrations," in *Conference on Robot Learning*, pp. 1930–1942, 2021.